REVIEW | Cores of Reproducibility in Physiology

CORP: Minimizing the chances of false positives and false negatives

Douglas Curran-Everett

Division of Biostatistics and Bioinformatics, National Jewish Health, Denver, Colorado; and Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, Denver, Colorado

Submitted 21 October 2016; accepted in final form 23 November 2016

Curran-Everett D. CORP: Minimizing the chances of false positives and false negatives. J Appl Physiol 122: 91–95, 2017. First published December 1, 2016; doi:10.1152/japplphysiol.00937.2016.—Statistics is essential to the process of scientific discovery. An inescapable tenet of statistics, however, is the notion of uncertainty which has reared its head within the arena of reproducibility of research. The Journal of Applied Physiology’s recent initiative, “Cores of Reproducibility in Physiology,” is designed to improve the reproducibility of research: each article is designed to elucidate the principles and nuances of using some piece of scientific equipment or some experimental technique so that other researchers can obtain reproducible results. But other researchers can use some piece of equipment or some technique with expert skill and still fail to replicate an experimental result if they neglect to consider the fundamental concepts of statistics of hypothesis testing and estimation and their inescapable connection to the reproducibility of research. If we want to improve the reproducibility of our research, then we want to minimize the chance that we get a false positive and—at the same time—we want to minimize the chance that we get a false negative. In this review I outline strategies to accomplish each of these things. These strategies are related intimately to fundamental concepts of statistics and the inherent uncertainty embedded in them.

estimation; hypothesis test; power; reproducibility; significance test

[Populations] always display variation ... .
—Sir Ronald A. Fisher (1925)

IT IS ONLY BECAUSE POPULATIONS always display variation that statistics is essential to the process of scientific discovery. An inescapable tenet of statistics, however, is the notion of uncertainty that has reared its head within the arena of reproducibility of research.

It just so happens that improving the reproducibility of research is also the focus of the Journal of Applied Physiology’s recent initiative, “Cores of Reproducibility in Physiology” (22). Each article in this initiative is designed to elucidate the principles and nuances of using some piece of scientific equipment or some experimental technique so that other researchers can obtain reproducible results (22). But other researchers can use some piece of equipment or some technique with expert skill and still fail to replicate a pivotal experimental result if they neglect to consider the fundamental concepts of statistics of hypothesis testing and estimation and their inescapable connection to the reproducibility of research.

If we want to improve the reproducibility of our research, then we want to minimize the chance that we get a false positive and—at the same time—we want to minimize the chance that we get a false negative. In this review I outline strategies to accomplish each of these things. These strategies are related intimately to fundamental concepts of statistics and the inherent uncertainty embedded in them. I begin with a brief overview of the null hypothesis, the foundation of any scientific experiment.

The Null Hypothesis: a Scientific Idea

When we design an experiment, we want to test a scientific idea. By tradition, this idea is called the null hypothesis (see Ref. 7). When we make an inference about a null hypothesis, we can make a mistake (Fig. 1). We can reject a true null hypothesis—we get a false positive—or we can fail to reject a false null hypothesis—we get a false negative.¹

We control the chance that we get a false positive when we define the critical significance level α, the probability that we reject the null hypothesis given that the null hypothesis is true. When we define α, we declare that we are willing to reject a true null hypothesis 100α% of the time.

In contrast, we control the chance that we get a false negative when we prescribe instead the chance that we get a true positive. We do this when we define power, the probability that we reject the null hypothesis given that the null hypothesis is false. Neyman and Pearson described a false positive as an error of the first kind and a false negative as an error of the second kind (see Ref. 7). Errors of the first kind and second kind are known also as type I and type II errors.

¹ Neyman and Pearson described a false positive as an error of the first kind and a false negative as an error of the second kind (see Ref. 7). Errors of the first kind and second kind are known also as type I and type II errors.
is false. In general, four things affect power: the critical significance level \( \alpha \), the standard deviation \( \sigma \) of the underlying population, the sample size \( n \), and the magnitude of the difference we want to be able to detect (Table 1; see also Ref. 8).

When the Null Hypothesis Is True: to Minimize False Positives

When we test a null hypothesis, rarely, if ever, do we expect that null hypothesis to be true (see Refs. 7 and 11). Rather, we usually know—at the very least, suspect—that our null hypothesis is not exactly true. If we happen to be mistaken about the validity of our null hypothesis, however, we would like to guard against an unsubstantiated discovery: we would like to avoid a false positive.

As I posited in Ref. 9, suppose we want to learn if some experimental treatment has a significant effect for which there is convincing statistical evidence but that is detectable only because of bias or error. We might ask if our two samples came from the same or different populations. This means we define the null and alternative hypotheses, \( H_0 \) and \( H_1 \), as

\[
H_0 : \text{The samples come from the same population.} \\
H_1 : \text{The samples come from the different populations.}
\]

If we want to know whether the populations have the same mean, we can write these as

\[
H_0 : \Delta \mu = 0 \\
H_1 : \Delta \mu \neq 0,
\]

where \( \Delta \mu \), the difference in population means, is the difference between the means of the treated and control populations.

In any experiment, regardless of the number of observations, the chance that we get a false positive depends only on the critical significance level \( \alpha \): we will obtain a false positive 100\( \alpha \)% of the time (Fig. 2).\(^2\) If we want to minimize the chance that we get a false positive, then we want \( \alpha \) to be more stringent than the traditional 0.05. A smaller, more stringent \( \alpha \) will increase the power of our test.

As I posited in Ref. 9, suppose we want to learn if some experimental treatment has a significant effect for which there is convincing statistical evidence but that is detectable only because of bias or error.

Values represent power for different combinations of sample size, critical significance level \( \alpha \), and effect size. Effect size is the difference between population means \( \mu_1 \) and \( \mu_0 \) expressed as a fraction of the population standard deviation \( \sigma \).\(^8\) An effect size of 0.5 means that \( \Delta \mu = 0.5 \sigma \), and an effect size of 1 means that \( \Delta \mu = \sigma \). If all else is equal, power increases when the sample size increases, the critical significance level \( \alpha \) increases, and the effect size increases.

- In general, four things affect power: the critical significance level \( \alpha \), the standard deviation \( \sigma \) of the underlying population, the sample size \( n \), and the magnitude of the difference we want to be able to detect (Table 1; see also Ref. 8).

Table 1. Power estimates for a two-sample t-test

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>Effect Size</th>
<th>Sample Size, ( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>0.01</td>
<td>0.25</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.29</td>
</tr>
<tr>
<td>0.05</td>
<td>0.25</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.56</td>
</tr>
<tr>
<td>0.10</td>
<td>0.25</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Values represent power for different combinations of sample size, critical significance level \( \alpha \), and effect size. Effect size is the difference between population means \( \mu_1 \) and \( \mu_0 \) expressed as a fraction of the population standard deviation \( \sigma \) (see Ref. 8). An effect size of 0.5 means that \( \Delta \mu = 0.5 \sigma \), and an effect size of 1 means that \( \Delta \mu = \sigma \). If all else is equal, power increases when the sample size increases, the critical significance level \( \alpha \) increases, and the effect size increases.

2 This refers to a false positive that occurs for statistical reasons. Bias or experimental error can result in a spurious true positive: there is an apparent effect for which there is convincing statistical evidence but that is detectable only because of bias or error.
only on the number of observations in our two groups (8).

Because we have defined our populations, power depends on the probability we reject our null hypothesis given it is false, is the chance of doing that if we design our experiment so that power, the probability we reject our null hypothesis if it is false. We can up our chances of thinking of this as the number of observations in our two groups (8).

It—almost—goes without saying that we would like to reject our null hypothesis if it is false. We can up our chances of doing that if we design our experiment so that power, the probability we reject our null hypothesis given it is false, is high. Because we have defined our populations, power depends only on the number of observations in our two groups (8).

Table 2. Probability a duplicate experiment will achieve P < 0.05

<table>
<thead>
<tr>
<th>$P_1$</th>
<th>$Pr (P_2 &lt; α = 0.05)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.40</td>
<td>0.13</td>
</tr>
<tr>
<td>0.30</td>
<td>0.18</td>
</tr>
<tr>
<td>0.20</td>
<td>0.25</td>
</tr>
<tr>
<td>0.10</td>
<td>0.38</td>
</tr>
<tr>
<td>0.05</td>
<td>0.50</td>
</tr>
<tr>
<td>0.025</td>
<td>0.61</td>
</tr>
<tr>
<td>0.01</td>
<td>0.73</td>
</tr>
<tr>
<td>0.005</td>
<td>0.80</td>
</tr>
<tr>
<td>0.001</td>
<td>0.91</td>
</tr>
<tr>
<td>0.00031</td>
<td>0.95</td>
</tr>
<tr>
<td>0.00002</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Values represent $P_1$, the $P$ value from an initial experiment, and $Pr (P_2 < α = 0.05)$, the probability that the $P$ value from a duplicate experiment will be less than $α = 0.05$. Ref. 9 discusses the calculation of these probabilities. See Fig. 4.

will also help us minimize the chance that we get a false negative when we attempt to reproduce the results of a pivotal, pioneering experiment.

When the Null Hypothesis Is False: to Minimize False Negatives

Now imagine we truly do know that our null hypothesis is false. Suppose we use the same null and alternative hypotheses we just did:

$$H_0 : \Delta \mu = 0$$

$$H_1 : \Delta \mu \neq 0.$$ 

If we define $\Delta \mu$, the difference in population means, to be 0.5 units (Fig. 3), then we know our null hypothesis is false.

It—almost—goes without saying that we would like to reject our null hypothesis if it is false. We can up our chances of doing that if we design our experiment so that power, the probability we reject our null hypothesis given it is false, is high. Because we have defined our populations, power depends only on the number of observations in our two groups (8).

Table 3. Percent inflated estimates of $\Delta \mu$ when $P < 0.05$

<table>
<thead>
<tr>
<th>$n$</th>
<th>Power</th>
<th>%Inflated</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.18</td>
<td>100</td>
</tr>
<tr>
<td>20</td>
<td>0.34</td>
<td>100</td>
</tr>
<tr>
<td>30</td>
<td>0.48</td>
<td>96</td>
</tr>
<tr>
<td>40</td>
<td>0.60</td>
<td>83</td>
</tr>
<tr>
<td>50</td>
<td>0.70</td>
<td>73</td>
</tr>
<tr>
<td>60</td>
<td>0.78</td>
<td>64</td>
</tr>
<tr>
<td>70</td>
<td>0.84</td>
<td>60</td>
</tr>
<tr>
<td>80</td>
<td>0.88</td>
<td>57</td>
</tr>
<tr>
<td>90</td>
<td>0.92</td>
<td>55</td>
</tr>
<tr>
<td>100</td>
<td>0.94</td>
<td>54</td>
</tr>
<tr>
<td>200</td>
<td>1.00</td>
<td>50</td>
</tr>
</tbody>
</table>

Values represent the results of 10,000 replications of a two-sample $t$-test in which the $n$ observations for each sample were drawn at random from the populations in Fig. 3. For each replication, we obtained a $P$ value with which to assess our null hypothesis $H_0 : \Delta \mu = 0$ and $\Delta \bar{y}$, the difference in sample means that estimates the true difference $\Delta \mu = 0.5$. Let us consider only those simulations for which we reject our null hypothesis: that is, for which $P < 0.05$. When power was 0.18 ($n = 10$), the estimated difference $\Delta \bar{y}$ exceeded the true difference $\Delta \mu$ in all those statistically significant simulations: 100% of the $\Delta \bar{y}$ values were inflated. In contrast, when power was 1.00 ($n = 200$), the estimated difference $\Delta \bar{y}$ exceeded the true difference $\Delta \mu$ in only half those simulations: 50% of the $\Delta \bar{y}$ values were inflated. After Ref. 9, Fig. 5.

With this background we are ready to explore the relationship of fundamental concepts of statistics—hypothesis testing and estimation—to reproducibility. With this background we are ready to explore the relationship of fundamental concepts of statistics—hypothesis testing and estimation—to reproducibility.

**Hypothesis testing.** Suppose we have a unique experimental result that is statistically unusual. Had we defined $α = 0.05$, this means our $P$ value is less than 0.05. If our null hypothesis is true, our result is unusual. We reject our null hypothesis. We have discovered something! We assume our result reflects a general phenomenon that other researchers will reproduce if they repeat our experiment. But will they?

If the $P$ value from our initial experiment is 0.05, then the probability a duplicate experiment will achieve $P < 0.05$—the probability it will achieve statistical significance—is 50% inflated. In contrast, when power was 1.00 ($n = 200$), the estimated difference $\Delta \bar{y}$ exceeded the true difference $\Delta \mu$ in only half those simulations: 50% of the $\Delta \bar{y}$ values were inflated. After Ref. 9, Fig. 5.

$^3$ The theoretical foundation of the next section is developed in—and was adapted from—Ref. 9.

Fig. 4. Probability a second experiment will result in $P < 0.05$ when the test statistic is $z$. If the critical significance level $α$ is 0.05, then the critical value $z^*$ is 1.96 (see Ref. 7). Suppose we conduct an experiment for which we compute an observed value $z_1$ that happens to be more extreme than $z^*$; that is, $z_1 > z^*$. This means $P_1$, the $P$ value associated with $z_1$, is less than the critical significance level $α$. This is unusual if the null hypothesis is true: we have a statistically significant result. If we assume the magnitude of the effect observed in this initial experiment equals the magnitude of the true effect, then the probability that the $P$ value from a second experiment will be less than 0.05 is depicted by the gray area. See Table 2. [Adapted with permission from Ref. 9.]
(Table 2). If the P value from our initial experiment is 0.01, then the probability a duplicate experiment will achieve $P < 0.05$ is about 75%. Only when the $P$ value from our initial experiment is 0.001 does the probability a duplicate experiment will achieve $P < 0.05$ exceed 90% (Fig. 4). Power, experimental design, and the actual test statistic (see Ref. 7) have little impact on this phenomenon (3).

*Estimation.* In contrast, power—with its inherent connection to the statistical benchmark of $\alpha$—does impact the reproducibility of point and interval estimates of the magnitude of some biologic effect (18). If an experiment of lower power happens to reject its null hypothesis—if the effect is statistically unusual—then our estimate of the magnitude of that effect will be exaggerated (9, 15); see Table 3 and Ref. 9. Experimental design and the actual test statistic have little impact on this phenomenon (18).

**The Impact of Multiple Comparisons**

When we do an experiment, we typically test more than one null hypothesis. This makes sense: an experiment can be costly in terms of time and resources, and we want to maximize what we learn. When we test more than one null hypothesis—when we have a family of comparisons—we are more likely to reject a true null hypothesis simply by virtue of the fact that we are testing more than one null hypothesis (Table 4). In this setting, we can minimize the chance that we get a false positive if we use a statistical procedure that controls for this phenomenon. The false discovery rate procedure has advantages over the Bonferroni procedure (1, 2, 5).

**Summary**

If we want to improve the reproducibility of our research, then we want to revamp how we apply the fundamental concepts of hypothesis testing and estimation to our science. This is how we can do that:

- When we design an experiment, estimate sample size so that power approaches 0.90.\(^5\) This minimizes the chance that we get a false negative and makes it less likely that we exaggerate the magnitude of the true effect.
- Define the critical significance level $\alpha$—the benchmark for how statistically unusual our result needs to be before we reject our null hypothesis—to be 0.005 or 0.001 (19, 20). This minimizes the chance that we get a false positive and makes it more likely a subsequent experiment will achieve $P < 0.05$.
- Power and $\alpha$ impact sample size. If we define $\alpha$ to be a more stringent 0.005 and if we design our experiment so that power approaches 0.90, might the sample size for our experiment be so large as to be practically impossible? Not necessarily (20).
- Think less about a simple $P$ value and more about the scientific importance of the confidence interval bounds for some experimental result (6, 10, 11, 14, 15, 20). Even a convincing $P$ value of 0.005 can be associated with a confidence interval whose bounds indicate an effect that is scientifically inconsequential: for example, [0.004, 0.01] (see Ref. 6).
- Treat with caution the potential magnitude of that experimental result (18).
- Control for multiple comparisons (5, 10). This minimizes the chance that we get a false positive simply because we make more than one comparison in a single experiment.
- Rely on repeated studies to accumulate evidence for some biological phenomenon (12, 17, 23). We can do this using simple inspection, as envisioned by Fisher (12, 13), or a formal meta-analysis (4, 16).

The notion of using repeated studies to accumulate evidence for some phenomenon is long established in science (12, 13, 21).\(^6\) In 1951 Yates (23) wrote

> Research workers...have to accustom themselves to the fact that in many branches of research the really critical experiment is rare, and that it is frequently necessary to combine the results of numbers of experiments dealing with the same issue in order to form a satisfactory picture of the true situation. . . . In such circumstances a number of experiments of moderate accuracy are of far greater value than a single experiment of very high accuracy.

In addition to the strategies listed above, perhaps it is time to also revisit that decades-old cornerstone of reproducibility.

**ACKNOWLEDGMENTS**

I dedicate this review to John Ludbrook (retired, Department of Surgery, The University of Melbourne, Melbourne, Victoria, Australia) who, in his own words, was a sometime academic surgeon, a sometime applied cardiovascular physiologist, and a biostatistician accredited by the Statistical Society of Australia. John and I have been corresponding since 2004 when he congratulated Dale Benos and me on the publication of our guidelines for reporting statistics. Over the years, on those rare occasions when he was able to tear himself away from watching cricket matches, John graciously reviewed some of the papers in my *Explorations in Statistics* series. My papers are better for it. And I have enjoyed tremendously our correspondence.

I thank Peter D. Wagner (University of California, San Diego, La Jolla, California) for his encouragement to write this review and for his helpful comments.

---

4 A sample mean $\bar{y}$ estimates some population mean $\mu$: a sample mean is a point estimate for some population mean. A confidence interval $[\bar{y} - a, \bar{y} + a]$, where $a$ is an allowance, is a range that we expect, with some level of confidence, to include the true value of a population parameter such as the mean: a confidence interval is an interval estimate for some population mean.

5 See Ref. 9 for some resources that can help with this.

6 See Ref. 9 for details.
DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author.

REFERENCES